REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
12/17/09	Final Report	10/01/2005 - 09/30/2009

4. TITLE AND SUBTITLE

Statistical Aspects of Information Integration

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-06-1-0096

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kolaczyk, Eric D.
Professor
Director, Program in Statistics
Department of Mathematics and Statistics
Boston University

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Trustees of Boston University
Office of Sponsored Programs
881 Commonwealth Avenue
Boston, MA 02215-2303

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR-251

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution unlimited.

13. SUPPLEMENTARY NOTES

None.

14. ABSTRACT

The focus of this basic research award, per the original proposal, was on the development of inference engines to be used within an overall information integration context, with emphasis on network-related environments. Two sub-areas received the primary attention in our work. The first was addressed successfully, while the second was pursued, found to be in need of modification, and the modification was pursued successfully. In addition, certain related issues regarding the acquisition of information from networks and its impact on inferential processes were successfully pursued as well. Overall, the research program was highly successful in achieving its stated goals, as well as in producing results on additional related goals that arose during the life of the award.

15. SUBJECT TERMS

Networks; statistics; anomaly detection; information integration.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			ONR
U	U	U		4	19b. TELEPHONE NUMBER (Include area code)

Final Report for N00014-06-1-0096

Title: Statistical Aspects of Information Integration

PI: Prof. Eric D. Kolaczyk
Department of Mathematics and Statistics
Boston University

Date: 12/17/2009

Overview

The focus of this basic research award, per the original proposal, was on the development of inference engines to be used within an overall information integration context, with emphasis on network-related environments. Two sub-areas received the primary attention in our work. The first was addressed successfully, while the second was pursued, found to be in need of modification, and the modification was pursued successfully. In addition, certain related issues regarding the acquisition of information from networks and its impact on inferential processes were successfully pursued as well. Overall, the research program was highly successful in achieving its stated goals, as well as in producing results on additional related goals that arose during the life of the award.

Specific Achievements

The first sub-area of research focus was on the use of non-parametric volume-based inferential strategies, in conjunction with dimension reduction techniques. We successfully developed methods to

1. extend the estimation of minimum volume sets from independent observations to dependent observations, characterizing the necessary mathematics and implementing the corresponding algorithms in software. [Di and Kolaczyk 2010]
2. augment the basic methodology so that it could be used as a testing/detection device, with corresponding control of false detection rates. [Scott and Kolaczyk 2007,2010]
3. use these minimum volume set methods for anomaly detection in moderate- to high-dimensional computer network traffic settings. [Chaabra et al 2008]

The second sub-area was the use of kernel fusion machines with numerous heterogeneous inputs, in conjunction with variable selection techniques and diagnostic tools. This thrust was pursued primarily within the context of large-scale biological databases, such as arise in computational biology, which were felt to be representative of many sources of network-based data more broadly. There it was found in initial studies that (i) the standard kernel methods and analogous probability-based methods performed similarly, and (ii) there were substantial sources of uncertainty in this type of data, for which work on extensions of probability-based methods would be much more likely to yield natural solutions. So the underlying technical machinery was shifted to integrative probability models, rather than integrative kernel methods. There we successfully developed methods to

1. integrate relational and hierarchical network data in a probabilistic framework that enforces coarse-to-fine hierarchical class relationships in classification, applied to protein function prediction in the context of the Gene Ontology network and protein interaction networks. [Jiang, Nariai, Steffen, Kasif, and Kolaczyk 2008]
2. integrate multiple sources of heterogeneous data types (i.e., both network and non-network, continuous and categorical), using naive-Bayes and conditional naive-Bayes techniques, and implement for the task of protein function prediction. [Nariai, Kolaczyk, and Kasif 2007; Jiang et al 2008; Jiang and Kolaczyk 2010]
3. incorporate into the network-based process prediction problem (inherent in #1 and 2 above) uncertainty information at the level of both network topology and data on the network-index process, yielding an ability to use inconsistencies between topology and process information to correct for such uncertainties. [Jiang, Gold, and Kolaczyk 2010]

Beyond the work in these two sub-areas, which constituted the main focus of our efforts, we also pursued a third line of research, focused on the problem of inferring association networks from temporally indexed data. Motivated by the application of needing to infer so-called `functional connectivity' networks in neuroscience, based on scalp-level voltage potential measurements, we

1. showed that statistical summaries of networks built by integrating measurements at multiple measurement sites provide potentially valuable predictive information on functional disabilities in epilepsy patients. [Kramer, Kolaczyk, and Kirsch 2008]
2. developed an inference and control procedure for creating such networks at multiple time points over a time period in a manner that maintains consistent levels of topological uncertainty throughout [Kramer, Eden, Cash, Kolaczyk 2009]

Finally, during the period of this award, the PI wrote and published a book in the general topic area of statistical analysis of network data [Kolaczyk 2009], which relies in part on many of the projects supported under both this ONR award and the PI's previous award with ONR. This book is the first of its kind, being a comprehensive survey of statistical methods for network analysis written, contrary to most discipline-specific treatments, from the perspective of the statistics itself. Organized according to a statistical taxonomy, the book covers both descriptive/exploratory and inferential statistical methods for network data. At the most recent Joint Statistical Meetings, held in Washington DC in 2009 and attended by over 6000 statisticians, the book was among the top ten sellers for Springer, the largest and one of the most prominent publishers in statistics.

Publications

Books

Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York, Springer.

Refereed Journal Articles

Nariai, N., Kolaczyk, E.D., and Kasif, S. (2007). Probabilistic model for protein function prediction from multiple types of genome-wide data. *PLoS ONE*, 2:3, e337.

Kramer, M.A., Kolaczyk, E.D., and Kirsch, H.E. (2008). Emergent network topology at seizure onset in humans. *Epilepsy Research*,79, 173-186.

Jiang, X., Nariai, N., Steffen, M., Kasif, S., and Kolaczyk, E.D. (2008). Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics*, 9:350.

Kolaczyk, E.D., Chua, D.B., and Barthelemy, M. (2009). Group-betweenness and co-betweenness: Inter-related notions of coalition centrality. *Social Networks*,31:3, 190-203.

Kramer, M.A., Eden, U.T., Cash, S.S., and Kolaczyk, E.D. (2009). Network inference -- with confidence -- from multivariate time series. *Physical Review E*, 79, 061916.

Scott, C. and Kolaczyk, E.D. (2010). Nonparametric assessment of contamination in multivariate data using minimum volume sets and FDR. *Journal of Computational and Graphical Statistics*, (to appear).

Di, J. and Kolaczyk, E.D. (2010). Complexity-penalized estimation of minimum volume sets for dependent data. *Journal of Multivariate Analysis*, (revised; under review).

Jiang, X., Gold, D.L., and Kolaczyk, E.D. (2009). Network-based auto-probit modeling for protein function

prediction. *Biometrics*, (revised; under review) .

Refereed Conference Proceedings

Scott, C.D. and Kolaczyk, E.D. (2007). Annotated minimum volume sets for nonparametric anomaly discovery. *Proceedings of the IEEE Statistical Signal Processing Workshop 2007*.

Chhabra, P., Scott, C., Kolaczyk, E.D., and Crovella, M. (2008). Distributed Spatial Anomaly Detection. *Proceedings of the IEEE Infocom 2008*.

Jiang, X., Nariai, N., Steffen, M., Kasif, S., Gold, D., and Kolaczyk, E.D. (2008). Combining Hierarchical Inference in Ontologies with Heterogeneous Data Sources Improves Gene Function Prediction. *Proceedings of the 2008 International IEEE Conference on Bioinformatics and Biomedicine*.

Book Chapters

Jiang, X. and Kolaczyk, E.D. (2010). Integration of network information for protein function prediction. In *Systems Biology and Signaling Networks*, Choi, S. (ed.). Springer. (To appear)

Invited Talks

"Path-based Sampling and Inference in the Internet: Implications of Network Structure." Classification Society of North America 2006 Meeting on Network Data Analysis and Data Mining. DIMACS Center, Rutgers University. May, 2006.

"Sampling Networks and the Inference of Network Characteristics". Network Science Workshop. Bloomington, Indiana. May, 2006.

"Distributed Spatial Anomaly Detection." DIMACS/DyDAn Workshop on Internet Tomography. Rutgers University. May, 2008.

"Whole-Network Methods for Traffic Analysis and Anomaly Detection." MITACS/MASCOS Joint Workshop on Fusion, Mining, and Security for Networks. McGill University, Canada. June, 2008.

"Statistical Multiresolution Analysis of Internet Traffic on Graphs: Good Idea of Wishful Thinking?" Workshop on Multiscale Representation, Analysis, and Modeling of Internet Data and Meausurements. IPAM, UCLA, Los Angeles. September, 2008.

"Network-based Auto-probit Modeling for Protein Function Prediction." Workshop on Network Modeling: Statistical Analysis of Network Data in Practice. Dublin, Ireland. June, 2009.

"Statistical Analysis of Network Data". Institut de Statistitique, l'Universite Catholique de Louvain, Belgium. Two-week short course. Sept/Oct 2009.

Additional invited presentations were made by the PI at various academic institutions over the period of this award, as well as contributed presentations by graduate students at various national meetings.

Graduate Students Supported

There were two graduate students supported on this award for an extensive period, and a third supported briefly near the end of the award. The first two wrote theses pertaining to each of the two main sub-topics studied under this award (described above). The last has recently begun work related to our other work on inference of association networks, and is now supported under another ONR award.

Mr. Jianing Di PhD, Boston University December, 2007
(Currently a statistician with Johnson & Johnson Research (San Diego).)

Ms. Xiaoyu Jiang PhD, Boston University December 2008
(Currently a statistician with Boehringer Ingelheim Pharmaceuticals (Connecticutt).)

Mr. Wes Viles Current PhD Student